JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-2, Issue-4 E-ISSN: 2347-2693

Developing New Software Metric Pattern Discovery for Text Mining

S.S. Patil[1*] and V.M. Gaikwad[2]

Department of Computer, Bharati Vidyapeeth University College of Engineering
Pune-411043, Maharashtra, India

www.ijcseonline.org

Received: 12/03/2014 Revised: 28/03/2014 Accepted: 27/04/2014 Published: 30/04/2014

Abstract— In this paper data mining technique available which useful in mining the pattern of text document. However efficiently Research of new pattern which is useful and update the discover patterns is still an open research issue, especially in text mining field. Existing text mining methods have term based approaches and they have problems of polysemy and synonymy. To reduce these problems people develop the phrase (pattern) based approach. The phrase based approach is better than the term based approach but many experiments are not done in phrase based approach. In this we present the new innovative and effective pattern discovery method which has a process of pattern deploying and pattern evolving, which improve the efficiency and updating new pattern for finding appropriate and interacting information.

Keywords-Text Mining, Pattern Mining, Pattern Evolving, Pattern Extraction

I. Introduction

We have develop the methods of knowledge discovery and data mining because in existing years there is quick growth in digital data, so it is needful to turn this data into useful information and knowledge. The applications like market analysis and business management can have advantages; by the use of this information and knowledge extracted from a large amount of data. Knowledge discovery is the process of nontrivial extraction of information from big or large databases. Information Is unknown and actually useful for users. Therefore Data mining is an important step in process of knowledge discovery in large databases.

In a last five to ten years, to perform different knowledge tasks the data mining technique are presented. These techniques are (I) association rule mining, (II) frequent item set mining, (III) sequential pattern mining, (IV) maximum pattern mining, and (V) closed pattern mining. Mostly focus of these techniques is efficient mining algorithms used to find particular pattern which is a reasonable and acceptable in time frame. Using these techniques, number of patterns for data mining is generated. How these patterns effectively use and update is still an open research issue. But there is still scope for the research. In this paper we have focus on the development of a knowledge discovery model which effectively use and update the patterns which are discovered and apply it to text mining.

Text mining is discovery which applied to text mining in the interesting knowledge in the text documents. It is a challenging issue to find exact knowledge (or features) in text documents to help users to find what they want? Firstly, information Retrieval (IR) provided many term-based methods to solve this challenge, such as Rocchio and probabilistic models, rough set models, BM25 and support vector machine (SVM) based filtering models. The profits of term- based methods include able computational performance as well as mature theories for term weight; they have emerged over the last twenty years from the IR and communities of machine learning. However, term- based methods suffer from the problems of synonymy and polysemy, where synonymy is multiple words having the same meaning and polysemy means a word has multiple meanings. The semantic meaning of many discovered terms is unsure for answering what users want. Over the years, people have often held the hypothesis that phrase-based approaches could perform better than the term-based ones, as phrases may carry more "semantics" like information. This hypothesis has not fared too well in the history of IR. Although phrases are less ambiguous and more discriminative than individual terms, the reasons for the discouraging performance as follow: 1) phrases have inferior statistical properties to terms, 2) they have low frequency of occasion, and 3) there are large numbers of excessive and noisy phrases among them.

In the presence of these setbacks, sequential patterns used in data mining community have turned out to be a promising alternative to phrases. Because sequential patterns enjoy good statistical properties like terms. To overcome the disadvantage of phrase-based approaches, pattern mining-based approaches is used, they adopted the concept of the closed sequential patterns and the pruned non closed patterns. These pattern mining based approaches have shown some extent improvements on the effectiveness.

There are two fundamental things regarding the effectiveness of pattern-based approach as low frequency and misinterpretation. Given a special topic, a highly frequent pattern is usually a general pattern, or a specific pattern of low frequency. If we not given the support, a lot of noisy

Corresponding Author: *S.S. Patil*

patterns would be searched. Misinterpretation means the measures used in pattern mining (e.g., “support” and “confidence”) which is not suitable for discovered patterns to give answer what we want. The main problem hence is how to use discovered patterns to accurately evaluate the weights of useful knowledge in text documents

Over the years, IR has developed many mature techniques which demonstrated the terms were which important features in text documents. However, general terms have larger weights (e.g., the term frequency and inverse document frequency (tf*idf) weighting scheme) because they can be frequently used in both relevant and irrelevant information. For example, term “LIB” may have larger weight than “JDK” in a certain of data collection; but we believe that term “JDK” is more specific than term “LIB” for describing “Java Programming Language”. Therefore, it is not adequate for evaluating the weights of the terms based on their classifications in documents for a given topic. These new method has been frequently used in developing information retrieval models.

In order to solve the above inappropriate statement, we present an effective pattern discovery technique, which evaluates discovered peculiarity of patterns. Then evaluate term weights according to the classification of terms in the discovered patterns rather than the distribution, for solving the wrong interpretation issue. It also considers the impact of patterns from the negative training to find ambiguous (noisy) patterns and try to reduce their influence for the frequency problem. The process of updating uncertain patterns can be referred as pattern evolution. The proposed approach can improve the correctness of evaluating term weights because discovered patterns are more definite than whole documents. We also conduct numerous experiments on the latest data collected, Reuters Corpus Volume 1 (RCV1) and Text Retrieval Conference (TREC) filtrating topics, to evaluate the proposed technique. The results show that the proposed technique outperforms up-to-date data mining-based methods, the state-of-the-art term- based methods and concept-based models.

II. RELATED WORK

The number of text representations has been developed in the past. A famous one is the bag of words that uses keywords (terms) as elements in the vector of the feature space. In, the *tf*idf* weighting systematic plan is used for text representation in the systematic Rocchio classifiers. With TFIDF, the universal IDF and entropy weighting plan is put forward and improves performance by an average of 30 percent. Different systematic plan for the bag of words represent technique. The problem of the bag of words technique is to select a limited number of features among an huge set of words in order to increase the system’s efficiency and avoid over fitting in order to reduce the number of features, many dimensionality reduction techniques are conducted by the use of feature selection techniques, such as Mutual Information, Information Gain, Odds ratio, Chi-Square, and so on.

The choice of a presentation depended on what is regards as the meaningful units of text and the meaningful natural language rules for the combination of these all units. With respect to the presentation of the content of documents, some research works not use individual words it uses phrases. In the combination of unigram and bigrams was selected for document indexing in Text Categorization (TC) and assess on a variety of feature evaluation functions (FEF). A phrase-based text representation was proposed for Web document management.

For text analysis the data mining techniques are used by extracting co-occurring terms as descriptive phrases from documents. However, the effectiveness of the text mining systems using phrases for text representation showed un-meaningful improvement. The reason behind that is, phrase-based method had “lower document frequency and lower consistency of assignment for terms.

Term-based ontology mining methods also provided some thoughts for text representations. For example, hierarchical clustering was used to determine synonymy and hyponymy relations in between keywords of text. Here pattern evolution technique was introduced in order to improve the fulfillment of term-based ontology mining. In data mining communities pattern mining has been extensively studied from many years. A variety of efficient algorithms such as Apriori-like algorithms, PrefixSpan , FP-tree, SPADE, SLPMiner, and GST have been put forward. For developing efficient mining algorithms for discovering patterns from a large data collection these research works are used. However, searching for useful and interesting patterns and rules was even now an open problem in the field of text mining, various text patterns are find using pattern mining techniques, such as sequential patterns, frequent item sets, multiple grams, and co-occurring terms, for building up a representation with these new types of characteristics. Nevertheless, how to effectively deal with big amount of discovered patterns is the challenging issue. For that challenging issue, closed sequential patterns have been used for text mining, which proposed that the potential for improving the performance of text mining the concept of closed patterns in text mining was useful. Pattern taxonomy model was developed and improved the effectiveness by using closed patterns in text mining.

Here the two-stage model is used to improve the performance of information filtering that is both term-based methods and pattern- based methods.

The modern computational technology that can help people to understand the meaning of text documents is Natural language processing (NLP). The NLP was struggling for dealing with uncertainties in person’s languages. Recently, a new concept-based model was presented to bridge the gap between NLP and text mining, which analyzed terms on the sentence and document levels. This model included three components. The first one analyzed the semantic structure of sentences; the second one constructed a conceptual ontological graph (COG) to describe the semantic structures; and the last one extracted top concepts based on the first two components to build characterized vectors using the standard vector space model. The advantage of the concept-based model is that it can effectively differentiate between non important terms and meaningful terms which define a sentence meaning. Compared with the above methods, the concept-based model depends upon its employed NLP techniques.

III. Proposed Work

As the existing system implemented evaluation of term support based up on their occurrence in documents, where as the proposed system is to extract the information what a user wants in the form of terms or patterns. The data set is loaded into the data base. The data set undergoes text pre-processing phase which includes tokenization, parts of speech, and word stemming and stop word removal methods. Then terms present in the data set are analyzed in order to find out which are positive documents and which are negative documents. Pattern taxonomy model, pattern deploying techniques were implemented on positive data set in next phase one after other to discover patterns. Simultaneously, negative sets undergo pattern evolving and shuffling techniques. Therefore the term supports are evaluated. Terms will be extracted from term extraction model. Time is evaluated in time evaluation module which displays how much time taken to do the above whole process.

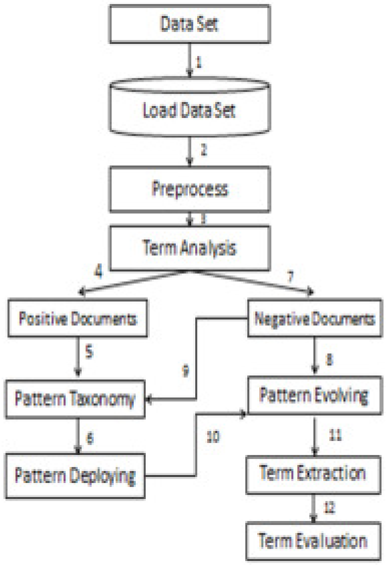


Fig1. Proposed Architecture Model Flow

A. *Datasets*

Dataset is collection of the data which present in tabular form. Here we use the data from database of big data which is the large and complex database on internet. That data is stored in big data in SGML format. SGML is the Slandered Generalized Markup Language.

IV. System Architecture

The proposed architecture is shown in Figure 1. This architecture shows the stepwise solution of our project. The basic step is to load SGML documents in our database. The next step is to convert that SGML document in XML format then remove stop word and text steaming. We removed this stop word and text steaming with the help of NLP (natural language process). Fig 1: System Architecture There is 5 sub modules of proposed system.

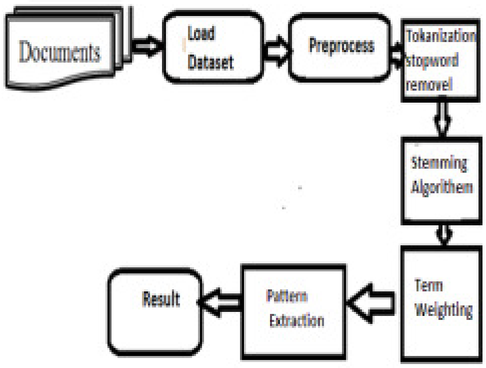


Fig 2: System Architecture

1) Loading documents

2) Text Preprocessing

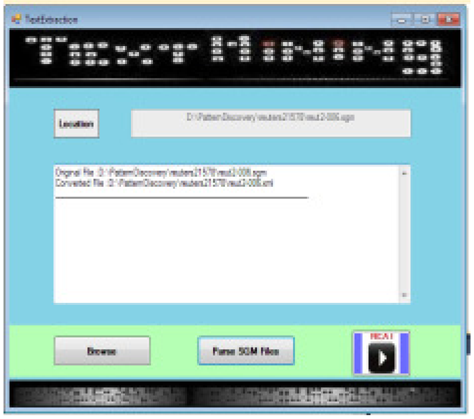
3) Term weighting

4) Dimensionally Reduction

5) Pattern Extraction

Loading documents

In this module, to load the SGML documents. The user retrieves one of documents. This document is given to next process that process is conversion of SGML to XML.

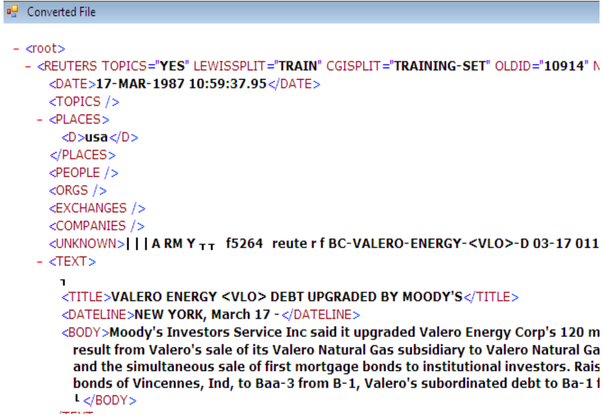


D:\PatternDiscovery\reuters21578\reut2-006.sgm

Orginal File :D:\PatternDiscovery\reuters21578\reut2-006.sgm
Converted File :D:\PatternDiscovery\reuters21578\reut2-006.xml

Fig.3 loading document

The converted document has the sub roots like topic, date, place, people etc.



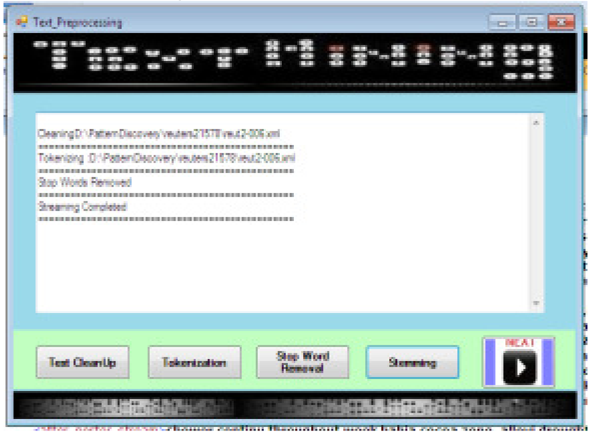
```
Converted File

- <root>
  - <REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="10914" N
      <DATE>17-MAR-1987 10:59:37.95</DATE>
      <TOPICS />
    - <PLACES>
        <D>usa</D>
      </PLACES>
      <PEOPLE />
      <ORGS />
      <EXCHANGES />
      <COMPANIES />
      <UNKNOWN>| | | A RM Y T T  f5264  reute r f BC-VALERO-ENERGY-<VLO>-D 03-17 011
    - <TEXT>
        ┐
        <TITLE>VALERO ENERGY <VLO> DEBT UPGRADED BY MOODY'S</TITLE>
        <DATELINE>NEW YORK, March 17 -</DATELINE>
        <BODY>Moody's Investors Service Inc said it upgraded Valero Energy Corp's 120 m
          result from Valero's sale of its Valero Natural Gas subsidiary to Valero Natural Ga
          and the simultaneous sale of first mortgage bonds to institutional investors. Rais
          bonds of Vincennes, Ind, to Baa-3 from B-1, Valero's subordinated debt to Ba-1 f
          └ </BODY>
```

Fig.4Converted SGML document to XML document

V. PRE-PROCESSING

The preprocessing phase of the study converts the original textual data in a data- mining-ready structure, where the most significant text-features that serve to differentiate between text-categories are identified, that means unstructured text documents are processed using natural language processing techniques to extract keywords labeling the items in that text documents. An effective preprocessor represents the document efficiently in terms of both space (for storing the document) and time (for processing retrieval requests) requirements and maintain good retrieval performance. The main objective of preprocessing is to obtain the key features or key terms from text documents and to enhance the relevancy between word and document and the relevancy between word and category. After this phase, classical data mining techniques are applied on the extracted data (keywords) to discover interesting patterns. Preprocessing technique contains

- Text Clean Up
- Tokenization
- Stop word removal
- Word stemming

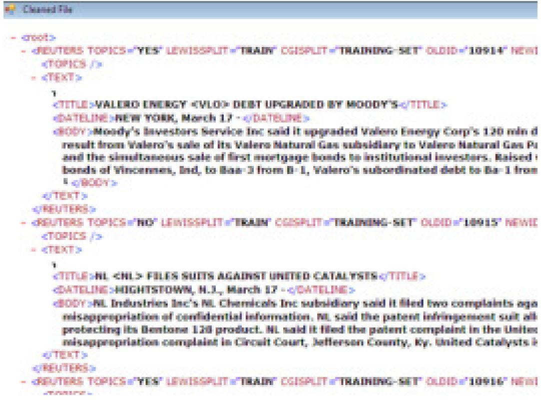


```
Cleaning D:\PatternDiscovery\reuters21578\reut2-006.xml
=================================================
Tokenizing D:\PatternDiscovery\reuters21578\reut2-006.xml
=================================================
Stop Words Removed
=================================================
Streaming Completed
=================================================
```

Fig.5 Pre-processing

A. *Text Clean Up*

After loading the document we see the XML document has number of sub roots like date, people, org, exchange, companies etc. In text cleanup process we will remove the roots that are unnecessary like people, org, exchange, people which is not having the meaning.

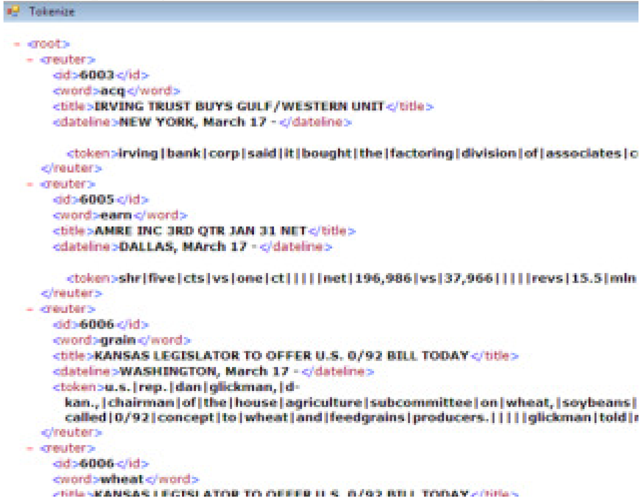


```
Cleaned File

- <root>
  - <REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="10914" NEWI
      <TOPICS />
    - <TEXT>
        ┐
        <TITLE>VALERO ENERGY <VLO> DEBT UPGRADED BY MOODY'S</TITLE>
        <DATELINE>NEW YORK, March 17 -</DATELINE>
        <BODY>Moody's Investors Service Inc said it upgraded Valero Energy Corp's 120 mln d
          result from Valero's sale of its Valero Natural Gas subsidiary to Valero Natural Gas Pa
          and the simultaneous sale of first mortgage bonds to institutional investors. Raised
          bonds of Vincennes, Ind, to Baa-3 from B-1, Valero's subordinated debt to Ba-1 from
          └ </BODY>
      </TEXT>
    </REUTERS>
  - <REUTERS TOPICS="NO" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="10915" NEWID
      <TOPICS />
    - <TEXT>
        ┐
        <TITLE>NL <NL> FILES SUITS AGAINST UNITED CATALYSTS</TITLE>
        <DATELINE>HIGHTSTOWN, N.J., March 17 -</DATELINE>
        <BODY>NL Industries Inc's NL Chemicals Inc subsidiary said it filed two complaints aga
          misappropriation of confidential information. NL said the patent infringement suit all
          protecting its Bentone 128 product. NL said it filed the patent complaint in the Unite
          misappropriation complaint in Circuit Court, Jefferson County, Ky. United Catalysts I
      </TEXT>
    </REUTERS>
  - <REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="10916" NEWI
```

Fig.6 Text Clean Up document

B. *Tokenization*

Given a character sequence and a defined document unit, tokenization is the task of hopping it up into pieces, called *tokens*, perhaps at the same time throwing away certain characters, such as punctuation. These tokens are often loosely referred to as terms or words. A *token* is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing.

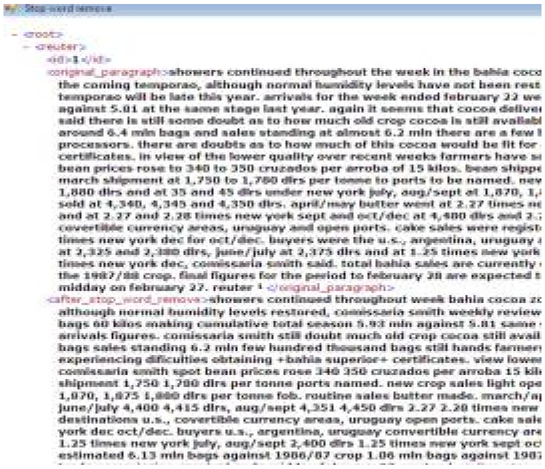


```
Tokenize

- <root>
  - <reuter>
      <id>6003</id>
      <word>acq</word>
      <title>IRVING TRUST BUYS GULF/WESTERN UNIT</title>
      <dateline>NEW YORK, March 17 -</dateline>

        <token>irving|bank|corp|said|it|bought|the|factoring|division|of|associates|c
    </reuter>
  - <reuter>
      <id>6005</id>
      <word>earn</word>
      <title>AMRE INC 3RD QTR JAN 31 NET</title>
      <dateline>DALLAS, MArch 17 -</dateline>

        <token>shr|five|cts|vs|one|ct|||||net|196,986|vs|37,966|||||revs|15.5|mln
    </reuter>
  - <reuter>
      <id>6006</id>
      <word>grain</word>
      <title>KANSAS LEGISLATOR TO OFFER U.S. 0/92 BILL TODAY</title>
      <dateline>WASHINGTON, March 17 -</dateline>
      <token>u.s.|rep.|dan|glickman,|d-
        kan.,|chairman|of|the|house|agriculture|subcommittee|on|wheat,|soybeans|
        called|0/92|concept|to|wheat|and|feedgrains|producers.|||||glickman|told|r
    </reuter>
  - <reuter>
      <id>6006</id>
      <word>wheat</word>
      <title>KANSAS LEGISLATOR TO OFFER U.S. 0/92 BILL TODAY</title>
```

Fig.7 Tokenization of document

C. *Stop Word Removal*

Many of the most frequently used words in English are useless in Information Retrieval (IR) and text mining. These words are called 'Stop words'. Stop-words, which are language-specific functional words, are frequent words that carry no information (i.e., pronouns, prepositions, conjunctions). Stop words are needed as it reduces indexing (or data) file size, Improve efficiency and effectiveness. Some stop words are a, the, after, to, too, from, before, some etc.



```
- <root>
  - <reuter>
      <id>1</id>
      <original_paragraph>showers continued throughout the week in the bahia coco
        the coming temporao, although normal humidity levels have not been rest
        temporao will be late this year. arrivals for the week ended february 22 we
        against 5.01 at the same stage last year. again it seems that cocoa delive
        said there is still some doubt as to how much old crop cocoa is still availabl
        around 6.4 mln bags and sales standing at almost 6.2 mln there are a few l
        processors. there are doubts as to how much of this cocoa would be fit for
        certificates. in view of the lower quality over recent weeks farmers have s
        bean prices rose to 340 to 350 cruzados per arroba of 15 kilos. bean shippe
        march shipment at 1,750 to 1,780 dlrs per tonne to ports to be named. new
        1,880 dlrs and at 35 and 45 dlrs under new york july, aug/sept at 1,870, 1,
        sold at 4,340, 4,345 and 4,350 dlrs. april/may butter went at 2.27 times ne
        and at 2.27 and 2.28 times new york sept and oct/dec at 4,480 dlrs and 2.
        covertible currency areas, uruguay and open ports. cake sales were regist
        times new york dec for oct/dec. buyers were the u.s., argentina, uruguay
        at 2,325 and 2,380 dlrs, june/july at 2,375 dlrs and at 1.25 times new york
        times new york dec, comissaria smith said. total bahia sales are currently
        the 1987/88 crop. final figures for the period to february 28 are expected t
        midday on february 27. reuter ¹ </original_paragraph>
      <after_stop_word_remove>showers continued throughout week bahia cocoa zo
        although normal humidity levels restored, comissaria smith weekly review
        bags 60 kilos making cumulative total season 5.93 mln against 5.81 same
        arrivals figures. comissaria smith still doubt much old crop cocoa still avail
        bags sales standing 6.2 mln few hundred thousand bags still hands farmer
        experiencing dificulties obtaining +bahia superior+ certificates. view lowe
        comissaria smith spot bean prices rose 340 350 cruzados per arroba 15 kil
        shipment 1,750 1,780 dlrs per tonne ports named. new crop sales light ope
        1,870, 1,875 1,880 dlrs per tonne fob. routine sales butter made. march/a
        june/july 4,400 4,415 dlrs, aug/sept 4,351 4,450 dlrs 2.27 2.28 times new
        destinations u.s., covertible currency areas, uruguay open ports. cake sal
        york dec oct/dec. buyers u.s., argentina, uruguay convertible currency are
        1.25 times new york july, aug/sept 2,400 dlrs 1.25 times new york sept oc
        estimated 6.13 mln bags against 1986/87 crop 1.06 mln bags against 198
```

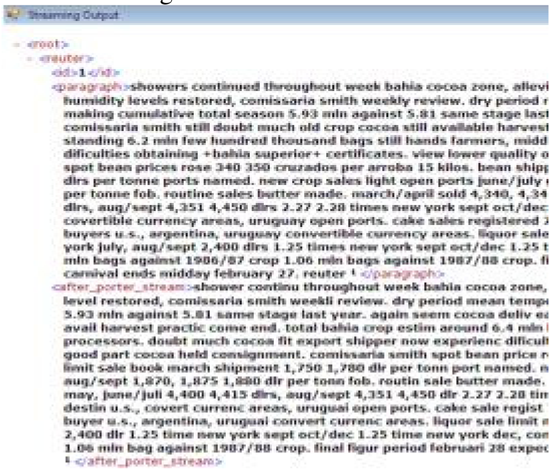
Fig.8 Stop Word Removal from document

D. Parts of Speech

A POS tagger marks the words in a text with labels corresponding to the part-of-speech of the word in that context. Part of speech tags the words according to grammatical context of words by dividing it into nouns, verbs and more. They are few typical tags for parts of speech they are NN (single noun), NNS (plural noun), VB (verb), VBD (verb, past tense), VBN (verb, past participle), IN (preposition), JJ (adjective), CC (conjunction, e.g., "and", "or"), PRP (pronoun) MD (modal auxiliary, e.g., "can", "will")

E. Word Stemming

Stemming techniques are used to find out the root/stem of a word. Stemming converts words to their stems, which incorporates a great deal of language-dependent linguistic knowledge. Behind stemming, the hypothesis is that words with the same stem or word root mostly describe same or relatively close concepts in text and so words can be conflated by using stems. The advantages of using stemming procedure is Stemming improves effectiveness of IR and text mining Matching similar words, mainly improve recall. It reduces indexing size, Combing words with same roots may reduce indexing size as much as 40-50%.



```
- <root>
  - <reuter>
      <id>1</id>
      <paragraph>showers continued throughout week bahia cocoa zone, allevi
        humidity levels restored, comissaria smith weekly review. dry period
        making cumulative total season 5.93 mln against 5.81 same stage last
        comissaria smith still doubt much old crop cocoa still available harvest
        standing 6.2 mln few hundred thousand bags still hands farmers, midd
        dificulties obtaining +bahia superior+ certificates. view lower quality o
        spot bean prices rose 340 350 cruzados per arroba 15 kilos. bean shipp
        dlrs per tonne ports named. new crop sales light open ports june/july
        per tonne fob. routine sales butter made. march/april sold 4,340, 4,34
        dlrs, aug/sept 4,351 4,450 dlrs 2.27 2.28 times new york sept oct/dec
        covertible currency areas, uruguay open ports. cake sales registered
        buyers u.s., argentina, uruguay convertible currency areas. liquor sale
        york july, aug/sept 2,400 dlrs 1.25 times new york sept oct/dec 1.25 t
        mln bags against 1986/87 crop 1.06 mln bags against 1987/88 crop. fi
        carnival ends midday february 27. reuter ¹ </paragraph>
      <after_porter_stream>shower continu throughout week bahia cocoa zone,
        level restored, comissaria smith weekli review. dry period mean temp
        5.93 mln against 5.81 same stage last year. again seem cocoa deliv e
        avail harvest practic come end. total bahia crop estim around 6.4 mln
        processors. doubt much cocoa fit export shipper now experienc dificul
        good part cocoa held consignment. comissaria smith spot bean price r
        limit sale book march shipment 1,750 1,780 dlr per tonn port named. n
        aug/sept 1,870, 1,875 1,880 dlr per tonn fob. routin sale butter made.
        may, june/juli 4,400 4,415 dlrs, aug/sept 4,351 4,450 dlr 2.27 2.28 tim
        destin u.s., covert currenc areas, uruguai open ports. cake sale regist
        buyer u.s., argentina, uruguai convert currenc areas. liquor sale limit
        2,400 dlr 1.25 time new york sept oct/dec 1.25 time new york dec, co
        1.06 mln bag against 1987/88 crop. final figur period februari 28 expec
        ¹ </after_porter_stream>
```

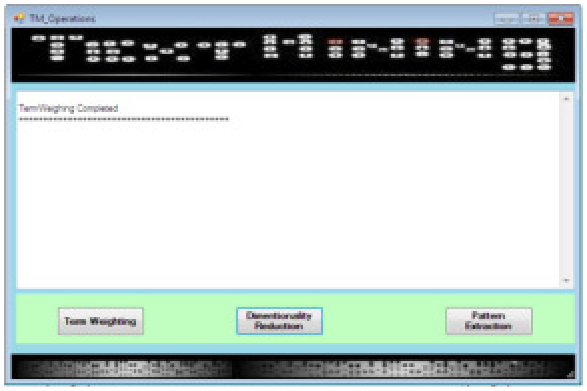
Fig.9 Result of Stemming Algorithm

VI. PATTERN TAXONAMY MODEL

In this phase, frequent patterns, sequential patterns, closed sequential patterns have been extracted. To improve the efficiency of the pattern taxonomy mining, an algorithm, SPMining, was proposed to find all closed sequential patterns, which used the well-known Apriori property in order to reduce the searching space. All the documents taken are split into paragraphs. So a given document d yields a set of paragraphs PS(d). Let D be a training set of documents, which consists of a set of positive documents, D+; and a set of negative documents, D-. Let T= {t1, t2 . . . tm} be a set of terms (or keywords) which can be extracted from the set of positive documents, D+. Given a termset X in document d, ˹X˺ is used to denote the covering set of X for *d*, which includes all paragraphs *dp*€ *PS(d)* such that X is subset of

dp, i.e., ˹X˺= {*dp/dp* € *PS(d)*,

X is subset of *dp}*. Frequent patterns are discovered in this phase using apriori algorithm in order to reduce the searching space for user. Frequent patterns are discovered based on absolute support and relative support.



```
TermWeighing Completed
**************************************************
```

Fig.10 Taxonomy Model Operation

Absolute support is the number of occurrences of X in *PS(d)* denoted by *supa*(X) = |˹x˺|. Relative support is the fraction of the paragraphs that contain the pattern denoted by *supr*(X) = |˹X˺|/|*PS(d)*|. A termset X is called frequent pattern if its supr (or supa) min sup, a minimum support. Table 1 lists a set of paragraphs for a given document d, where *PS(d)={dp1,dp2,…dp6},*and duplicate terms were removed.

Table 1:A Set Of Paragraphs

Paragraph	Terms
dp1	*t1 t2*
dp2	*t3 t4 t6*
dp3	*t3 t4 t5 t6*
dp4	*t3 t4 t5 t6*
dp5	*t1 t2 t6 t7*
dp6	*t1 t2 t6 t7*

Let min sup = 50%, ten frequent patterns are obtained in Table 1 using the above definitions. Table 2 illustrates the ten frequent patterns and their covering sets.

Table 2: Frequent Patterns And Covering Sets

Frequent patterns	Covering sets
{t3, t4, t6}	*{dp2, dp3, dp4}*
{t3,t4}	*{ dp2, dp3, dp4}*
{ t3, t6}	*{ dp2, dp3, dp4}*
{t4, t6}	*{ dp2, dp3, dp4}*
{t3}	*dp2, dp3, dp4}*
{t4}	*{ dp2, dp3, dp4}*
{t1, t2}	*{dp1, dp5, dp6}*
{t1}	*{ dp1, dp5, dp6}*
{t2}	*{ dp1, dp5, dp6}*
{t6}	*{dp2, dp3, dp4, dp5}*

Not all frequent patterns in Table 2 are useful. For example, pattern {*t3, t4} always* occurs with term *t6* in paragraphs, i.e., the shorter pattern, *{t3, t4}*, is always a part of the larger pattern, *{t3, t4, t6}*, in all of the paragraphs. Hence, the shorter one, *{t3; t4}*, is a noise pattern and expect to keep the larger pattern, *{t3, t4, t6}*, only.

Given a termset X, its covering set ˹X˺ is a subset of paragraphs. Similarly, given a set of paragraphs Y is a subset of *PS(d)*, Its termset can be defined, which satisfies *Termset*(Y) = {t| *dp Y => t dp}* The closure of X is defined as follows:
Cls(X) = termset(˹X˺).
A pattern X (also a termset) is called closed if and only if X = Cls(X).
Let X be a closed pattern. We can prove that
Supa(X1) < supa(X),
For all patterns X is a subset of X1; otherwise, if supa(X1) = supa(X), we have *˹X1˺ = ˹X˺*
Where *supa(X1) and* supa*(X)* are the absolute support of pattern *X1* and *X*, respectively.

Pattern taxonomy has been evaluated to discover closed patterns. Patterns can be structured into a taxonomy by using the (or subset) relation. For the example of Table 1, where a set of paragraphs of a document are illustrated, and the discovered 10 frequent patterns in Table 2 if assuming *min_sup =50%.* There are, however, only three closed patterns in this example. They are {*t3, t4, t6}, {t1, t2},* and *{t6}*. Fig 1 illustrates an example of the pattern taxonomy for the frequent patterns in Table 2, where the nodes represent frequent patterns and their covering sets; non-closed patterns can be pruned; the edges are "is-a" relation. After pruning, some direct "is-a" retaliations may be changed, for example, pattern *{t6}* would become a direct sub pattern of *{t3, t4, t6}* after pruning non-closed patterns. From frequent patterns and closed patterns, closed sequential patterns have been discovered using SP Mining algorithm.

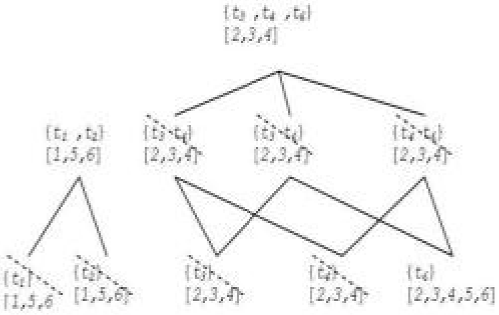


Fig. 11 Pattern Taxonomy

A. Term-weighting
There are many ways to define the term-weighting for the nonzero entries in such a vector. For example, we can simply set *TF*(*d*; *t*) = 1 if the term *t* occurs in the document *d*, or use the term frequency *freq*(*d*; *t*), or the relative term frequency, that is, the term frequency versus the total number of occurrences of all the terms in the document. There are also other ways to normalize the term frequency. For example, the Cornell SMART system uses the following formula to compute the (normalized) term frequency:

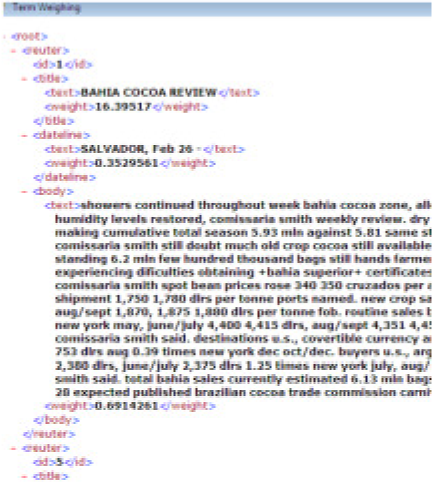
$$TF(d,t) = \begin{cases} 0 & if\,freq(d,t) = 0 \\ 1 + \log(1 + \log(freq(d,t))) & otherwise. \end{cases}$$

Besides the term frequency measure, there is another important measure, called inverse document frequency (IDF), that represents the scaling factor, or the importance, of a term *t*. If a term *t* occurs in many documents, its importance will be scaled down due to its reduced discriminative power. For example, the term *database systems* may likely be less important if it occurs in many research papers in a database system conference.

According to the same Cornell SMART system, *IDF*(*t*) is defined by the following formula:

$$|IDF(t) = \log \frac{1 + |d|}{|d_t|},$$

where *d* is the document collection, and *dt* is the set of documents containing term *t*.

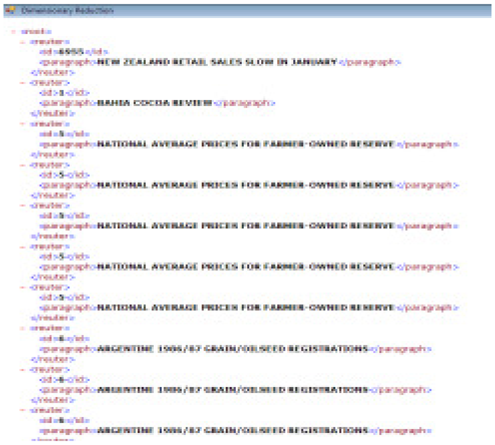


```
- <root>
  - <reuter>
      <id>1</id>
    - <title>
        <text>BAHIA COCOA REVIEW</text>
        <weight>16.39517</weight>
      </title>
    - <dateline>
        <text>SALVADOR, Feb 26 - </text>
        <weight>0.3529561</weight>
      </dateline>
    - <body>
        <text>showers continued throughout week bahia cocoa zone, all
          humidity levels restored, comissaria smith weekly review. dry
          making cumulative total season 5.93 mln against 5.81 same st
          comissaria smith still doubt much old crop cocoa still available
          standing 6.2 mln few hundred thousand bags still hands farme
          experiencing dificulties obtaining +bahia superior+ certificates
          comissaria smith spot bean prices rose 340 350 cruzados per a
          shipment 1,750 1,780 dlrs per tonne ports named. new crop sa
          aug/sept 1,870, 1,875 1,880 dlrs per tonne fob. routine sales b
          new york may, june/july 4,400 4,415 dlrs, aug/sept 4,351 4,45
          comissaria smith said. destinations u.s., covertible currency a
          753 dlrs aug 0.39 times new york dec oct/dec. buyers u.s., arg
          2,380 dlrs, june/july 2,375 dlrs 1.25 times new york july, aug/
          smith said. total bahia sales currently estimated 6.13 mln bags
          28 expected published brazilian cocoa trade commission carni
        <weight>0.6914261</weight>
      </body>
    </reuter>
  - <reuter>
      <id>5</id>
    - <title>
```

Fig.12 Term Weighting

B. Dimensionally Reduction

In previous term weighting we have the weights for every term or the token. Here in dimensionally reduction we will remove the terms or tokens who have weights less than user define and which is more than 0.



```
- <root>
  - <reuter>
      <id>6555</id>
      <paragraph>NEW ZEALAND RETAIL SALES SLOW IN JANUARY</paragraph>
    </reuter>
  - <reuter>
      <id>1</id>
      <paragraph>BAHIA COCOA REVIEW</paragraph>
    </reuter>
  - <reuter>
      <id>5</id>
      <paragraph>NATIONAL AVERAGE PRICES FOR FARMER-OWNED RESERVE</paragraph>
    </reuter>
  - <reuter>
      <id>5</id>
      <paragraph>NATIONAL AVERAGE PRICES FOR FARMER-OWNED RESERVE</paragraph>
    </reuter>
  - <reuter>
      <id>5</id>
      <paragraph>NATIONAL AVERAGE PRICES FOR FARMER-OWNED RESERVE</paragraph>
    </reuter>
  - <reuter>
      <id>5</id>
      <paragraph>NATIONAL AVERAGE PRICES FOR FARMER-OWNED RESERVE</paragraph>
    </reuter>
  - <reuter>
      <id>5</id>
      <paragraph>NATIONAL AVERAGE PRICES FOR FARMER-OWNED RESERVE</paragraph>
    </reuter>
  - <reuter>
      <id>6</id>
      <paragraph>ARGENTINE 1986/87 GRAIN/OILSEED REGISTRATIONS</paragraph>
    </reuter>
  - <reuter>
      <id>6</id>
      <paragraph>ARGENTINE 1986/87 GRAIN/OILSEED REGISTRATIONS</paragraph>
    </reuter>
  - <reuter>
      <id>6</id>
      <paragraph>ARGENTINE 1986/87 GRAIN/OILSEED REGISTRATIONS</paragraph>
```

Fig.13Dimensionally Reduction

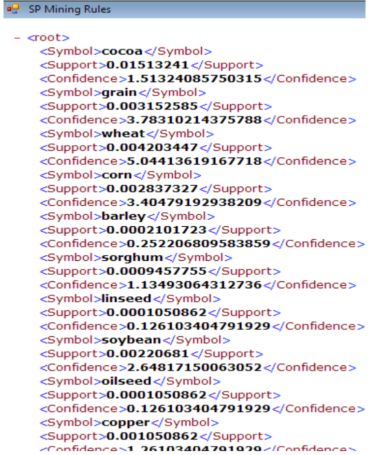
C. Pattern Extraction

The searching process undergoes in frequent patterns, closed and closed sequential patterns in order to extract the information which a user wants in form of terms. Along with terms, *nij* is its support in di which is the total absolute paragraph (D+ or D-) is extracted which contains supports given by closed patterns that contain *tij*; or *nij* is total number of closed patterns that contain *tij*. The process of calculating d-patterns can be easily described using pattern taxonomy model algorithm. If the term what is required has found out during searching process in any of the extracted patterns, searching is aborted. Here for the support and confidence following formulas are used

$$support(X \Rightarrow Y) = P(X \cup Y).$$

$$confidence(X \Rightarrow Y) = P(Y|X).$$

Here support indicates that a transaction contains both *X* and *Y*, that is, the union of itemsets *X* and *Y* and confidence indicates conditional probability that is, the probability that a transaction containing *X* also contains *Y*.



```
- <root>
    <Symbol>cocoa</Symbol>
    <Support>0.01513241</Support>
    <Confidence>1.51324085750315</Confidence>
    <Symbol>grain</Symbol>
    <Support>0.003152585</Support>
    <Confidence>3.78310214375788</Confidence>
    <Symbol>wheat</Symbol>
    <Support>0.004203447</Support>
    <Confidence>5.04413619167718</Confidence>
    <Symbol>corn</Symbol>
    <Support>0.002837327</Support>
    <Confidence>3.40479192938209</Confidence>
    <Symbol>barley</Symbol>
    <Support>0.0002101723</Support>
    <Confidence>0.252206809583859</Confidence>
    <Symbol>sorghum</Symbol>
    <Support>0.0009457755</Support>
    <Confidence>1.13493064312736</Confidence>
    <Symbol>linseed</Symbol>
    <Support>0.0001050862</Support>
    <Confidence>0.126103404791929</Confidence>
    <Symbol>soybean</Symbol>
    <Support>0.00220681</Support>
    <Confidence>2.64817150063052</Confidence>
    <Symbol>oilseed</Symbol>
    <Support>0.0001050862</Support>
    <Confidence>0.126103404791929</Confidence>
    <Symbol>copper</Symbol>
    <Support>0.001050862</Support>
    <Confidence>1.26103404791929</Confidence>
```

Fig.14 Pattern Extraction

VII. EXPERIMENTAL METHOD

In the experimental procedure, few users will be allowed to choose the particular document what they want. In search option of this application developed, users enter the required keyword of their interest, such that selected document undergoes preprocessing, terms are analyzed, frequent patterns, closed and closed sequential patterns are extracted. If the required key word is obtained in any one of extracted patterns, searching process will be aborted. All these results will be analyzed along with the feedback taken from every user to know the accuracy of process. The time taken for obtaining the optimist result will be compared with time taken for term based approach. Hence it can be proved that proposed approach is more accurate than term based approach.

VIII. CONCLUSION

This paper is intended to extract the useful information for users what he want. The extracted information will be in form of terms. In preprocessing, the document is tokenized, word stemming is done, and stop words are removed. The terms are analyzed, which undergoes pattern taxonomy, pattern deployment, pattern evolving, and term extraction to get accurate result.

References

[1]. T. Rose, M. Stevenson, and M. Whitehead, "The Reuters Corpus Volume1—From Yesterday's News to Today's Language Resources," Proc. Third Int'l Conf. Language Resources and Evaluation, pp. 29-31, 2002.

[2]. M.F. Porter, "An Algorithm for Suffix Stripping," Program, vol. 14, no. 3, pp. 130-137, 1980.

[3]. S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining" Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.

[4]. Y. Li and N. Zhong, "Mining Ontology for Automatically Acquiring Web User Information Needs," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 4, pp. 554-568,Apr. 2006.

[5]. Efficient Pattern Discovery for Text Mining, Ning Zhong, Yuefeng Li, Sheng-Tang Wu. IEEE Transactions on Knowledge and Data Engineering, VOL. 24, NO. 1, JANUARY 2012

[6]. Text Mining Approaches to Extract Interesting Association Rules from Text Documents Authors: Vishwadeepak Singh Baghela, Dr.S.Tripathi.

[7]. Evaluating Preprocessing Techniques in Text Categorization. Authors: V. Srividhya, R.

Authors Profile

Santosh S. Patil doing M.Tech from Department of Computer, Bharati Vidyapeeth University College Of Engineering, Pune-411043, Maharashtra .India.*sp070384@gmail.com*

Associate Prof. Mrs. V.M. Gaikwad
Department of Computer, Bharati Vidyapeeth University College Of Engineering, Pune-411043, Maharashtra *.India.sp070384@gmail.com*